BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 10: e184100
doi: 10.3897/biss.10.184100

OPEN ACCESS

Conference Abstract

Towards Building a National Genetic Reference Library for Species Identification in Support of Biodiversity Research and Public Policies in France

Liliana Ballesteros Mejia‡, Nirmala Seon-Massin‡, Gargominy Olivier§, Aurélie Lacoeuilhe§, Rodolphe Rougerie|

‡ Direction de l'Expertise - DGD-REVE. Muséum national d'Histoire naturelle, Paris, France
§ PatriNat (OFB, MNHN, CNRS, IRD), Paris, France
| Institut de Systématique, Évolution, Biodiversité (ISYEB, UMR7205), Muséum national d'Histoire naturelle, CNRS, EPHE-PSL, Université des Antilles, Sorbonne Université, Paris, France

Corresponding author: Liliana Ballesteros Mejia (ballesteros.liliana@gmail.com)

Received: 02 Jan 2026 | Published: 17 Feb 2026

Citation: Ballesteros Mejia L, Seon-Massin N, Olivier G, Lacoeuilhe A, Rougerie R (2026) Towards Building a National Genetic Reference Library for Species Identification in Support of Biodiversity Research and Public Policies in France. Biodiversity Information Science and Standards 10: e184100.
https://doi.org/10.3897/biss.10.184100

Abstract

Biological diversity is being strongly affected by the rapid environmental changes of the Anthropocene. Our ability to identify species presents a major challenge for monitoring the impacts of these changes and for designing sound public policies in response. Molecular methods, such as DNA barcoding, metabarcoding, and environmental DNA (eDNA) offer promising solutions; however, the effectiveness of these approaches relies heavily on the completeness and reliability of DNA reference libraries.

To address this challenge, and as part of France's national biodiversity strategy for 2030, the National Museum of Natural History in Paris (MNHN) and the French Biodiversity Agency (OFB) are developing the national reference of genetic sequences, called GenRef. This public and open-access tool will provide high-quality DNA sequences linked to reliable taxonomic identifications of species, and where available, to voucher

© Ballesteros Mejia L et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

specimens in biological collections, for taxa present in mainland France and its overseas territories (Lacoeuilhe et al. 2025).

GenRef will centralise, structure, and validate genetic reference sequences and their associated metadata using semi-automated workflows and expert review. Closely integrated with the France's national nomenclatural and taxonomic registry (TAXREF, Gargominy 2025), it will be interoperable with major international databases housing sequences such as European Nucleotide Archive (ENA), Barcode of Life Data Systems (BOLD), National Center for Biotechnology Information (NCBI GenBank), Unite (for fungal ITS sequences). In addition it will collaborate with French research institutions and natural history collections to incorporate their data and expertise, while promoting the acquisition of new data.

The GenRef workflow is organised into seven phases throughout which we can see the different stages of a GenRef Dataset.

1. **Data acquisition:** selection of taxon names in TAXREF (Gargominy 2025) and desired genetic marker(s), followed by automatic querying and retrieval of genetic sequences and associated metadata from major international genetic databases (e.g., ENA, BOLD, GenBank, Unite) via APIs. This step results in an initial dataset for GenRef (a working dataset for subsequenting processing, not a formal entity).
2. **Data filtering**: verification of mandatory fields and optional removal of contaminant sequences, such as those originating from non-target organisms (e.g., human, cow, mouse) or symbionts like *Wolbachia*. The output is a filtered version of the working GenRef dataset.
3. **Data extension:** addition and population of fields required for score calculations using previously compiled information. Examples of such fields may include monophyly status, average genetic distance to sequences sharing the same species name (derived from Neighbor-Joining (NJ) tree and distance matrix calculations), sequence length and percentage of ambigous base pairs. Additional metadata fields may include biogeographic status, taxonomic validity, identification by a recognized expert and specimen collection status. This step results in an extended working dataset used for downstream evaluation.
4. **Scoring and classification:** calculation of two scores for each record (i) a sequence identification reliability score based on sequence quality and identification accuracy, and (ii) completeness and relevance score based on the metadata completeness and provenance. The scoring framework is designed to be transparent, internally consistent, and adaptable as GenRef evolves, allowing records to be ranked into quality tiers for subsequent validation.
5. **Expert validation:** domain experts review the compiled molecular, taxonomic and metadata evidence for each record and use tools provided in the interface to confirm or adjust the automatic classifications. Their input generates an expert score and ranking, ensuring that expert judgement is integrated into the process rather than serving as a mere approval step.
6. **Moderation and arbitration:** resolves potential conflicts between automatic and expert classifications. This phase aggregates the expert ratings and applies

moderation rules to produce a single consensus score for each sequence. The implementation may involve a combination of algorithm checks and human oversight (final design is under development).

7. **User Interface:** a dedicated web interface provides access to sequences, with flexible options for data selection, filtering and visualisation, enabling researchers and stakeholders to explore and download curated data.

Ultimately, GenRef will provide access to sequences through a web-based user interface generating on-demand reference sequence libraries that can be used for different applications. These will range from the simple identification of individual specimens to surveillance of biological communities through eDNA analysis to support efforts in biodiversity knowledge, monitoring, research, conservation and policy making in France.

Keywords

eDNA, GenRef, French referential of genetic sequences, expert-validated DNA sequences, DNA sequence quality assurance.

Presenting author

Liliana Ballesteros-Mejia

Presented at

Living Data 2025

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Gargominy O (2025) TAXREF Version 4.17. Checklist dataset. In: PatriNat (Ed.) UAR PatriNat (OFB-MNHN-CNRS-IRD). [In English]. https://doi.org/10.15468/vqueam
- Lacoeuilhe A, Denys G, Dupont P, Archambeau A, Bed'How B, Bouix T, Brisset J, Cammas M, Detaï A, Dusoulier F, Gachon C, Gargominy O, Gaudeul M, Haÿ V, Herard K, Bras YL, Gall LL, Maurel N, Mazancourt Vd, Mennesson M, Pamerlon S, Poncet R, Puillandre N, Rougerie R, Samadi S, Sire L, Vinet C, Poncet L, Touroult J, et al. (2025) Référentiel des séquences génétiques des espèces de France : note pour la mise en place d'un nouvel outil national. In: PatriNat (Ed.) (OFB-MNHN-CNRS-IRD). URL: https://hal.science/hal-04931482v1